CS 58500 – Theoretical Computer Science Toolkit

Lecture 4 (01/29)

Concentration Inequality III

https://ruizhezhang.com/course_spring_2026.html

Recap

For a random variable Z, define its log-moment generating function $\psi(\theta) \coloneqq \log \mathbb{E}\left[e^{\theta(Z-\mathbb{E}[Z])}\right]$.

$$\Pr[Z - \mathbb{E}[Z] \geq t] \leq \exp\left(\inf_{\theta \geq 0} -\theta t + \psi(\theta)\right)$$

Let $X_1, \dots, X_n$ be independent random variables and $Z \coloneqq X_1 + \cdots + X_n$

- Hoeffding's inequality: if $a_i \leq X_i \leq b_i$ for $i \in [n]$, then

$$\Pr[|Z - \mathbb{E}[Z]| \geq t] \leq 2\exp\left(-\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)$$

- Chernoff bound: if X_i's are Bernoulli random variables, then

$$\Pr\big[|Z - \mathbb{E}[Z]| \geq t\mathbb{E}[Z]\big] \leq 2\exp(-t^2\mathbb{E}[Z]/3)$$

- Bernstein's inequality: if $|X_i - \mathbb{E}[X_i]| \leq b$ for $i \in [n]$, then

$$\Pr[|Z - \mathbb{E}[Z]| \geq t] \leq 2\exp\left(-\frac{t^2/2}{\mathrm{Var}[Z] + bt/3}\right)$$

Today's Lecture

- Tensorization of Variance (Revisited)
- Azuma-Hoeffding Inequality
- Applications
  - Pattern Matching
  - Learning Theory and Glivenko-Cantelli Theorem

Tensorization of Variance (Revisited)

Theorem. Suppose $X_1, \dots, X_n$ are independent random variables. Let $Z = f(X_1, \dots, X_n)$. Then

$$\mathrm{Var}[Z] \leq \mathbb{E}\left[\sum_{i=1}^{n} \mathrm{Var}_i[Z]\right]$$

where $\mathrm{Var}_i[Z](x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n) \coloneqq \mathrm{Var}[f(x_1, \dots, x_{i-1}, X_i, x_{i+1}, \dots, x_n)]$

Detour: Conditional Expectation

Conditional expectation from introductory probability class

- Let X be a random variable with $\mathbb{E}[|X|] < \infty$, and Y be another random variable with $\Pr[Y = y] > 0$
- Then we can define $\mathbb{E}[X|Y = y] = \sum_x x \Pr[X = x, Y = y]/\Pr[Y = y]$
- $Z = \mathbb{E}[X|Y]$ is a random variable such that $\Pr\big[Z = \mathbb{E}[X|Y = y]\big] = \Pr[Y = y]$

Issue: consider a 2-d Gaussian $(X, Y) \sim \mathcal{N}(0, \Sigma)$ with probability density function $g(x, y)$. What is $\mathbb{E}[X|Y = y]$? Intuitively, it is natural to define it as

$$\mathbb{E}[X|Y = y] = \frac{\int x g(x, y) \mathrm{d}x}{\int g(x, y) \mathrm{d}x}$$

However, for any $y \in \mathbb{R}$, $\Pr[Y = y] = 0$!

- We need measure-theoretic probability theory, where $\mathbb{E}[X|Y]$ is directly defined as a random variable (instead of for each $Y = y$) satisfying $\mathbb{E}[\mathbb{E}[X|Y]h(Y)] = \mathbb{E}[Xh(Y)]$ for any test function h

Detour: Conditional Expectation

Useful properties of conditional expectation

- Tower property:

$$\mathbb{E}[\mathbb{E}[X|Y]|Y,Z] = \mathbb{E}[X|Y] = \mathbb{E}[\mathbb{E}[X|Y,Z]|Y]$$

-

$$\mathbb{E}[X] = \mathbb{E}\big[\mathbb{E}[X|Y]\big]$$

-

$$\mathbb{E}[XY|X,Z] = X\mathbb{E}[Y|X,Z]$$

- For any invertible function f,

$$\mathbb{E}[X|Y] = \mathbb{E}[X|f(Y)]$$

Detour: Martingale

A sequence of random variables $Z_1, Z_2, \ldots$ is a martingale with respect to sequence $X_1, X_2, \ldots$ if for all $i \geq 0$,

- Z_i is a function of $X_1, \ldots, X_i$
- $\mathbb{E}[|Z_i|] < \infty$
- $\mathbb{E}[Z_{i+1} | X_1, \ldots, X_i] = Z_i$

In particular, we say $Z_1, Z_2, \ldots$ is a martingale if it's a martingale with respect to itself.

Example: Gambling

- Suppose a gambler places bets on a sequence of fair games: bets can increase/decrease based on history
- Let X_t be amount he wins at step t (could be negative)
- Let $Z_t \coloneqq \sum_{i \in [t]} X_i$ be total winning at end of t-th step
- $Z_1, Z_2, \ldots$ is a martingale, since $\mathbb{E}[Z_{i+1} | X_1, \ldots, X_i] = Z_i + \mathbb{E}[X_{i+1}] = Z_i$

Detour: Martingale

A sequence of random variables $Z_1, Z_2, \ldots$ is a martingale with respect to sequence $X_1, X_2, \ldots$ if for all $i \geq 0$,

- Z_i is a function of $X_1, \ldots, X_i$
- $\mathbb{E}[|Z_i|] < \infty$
- $\mathbb{E}[Z_{i+1} | X_1, \ldots, X_i] = Z_i$

In particular, we say $Z_1, Z_2, \ldots$ is a martingale if it's a martingale with respect to itself.

Lemma. Let $Z_1, Z_2, \ldots$ be a martingale with respect to $X_1, X_2, \ldots$. Then,

$$\mathbb{E}[Z_n] = \mathbb{E}[Z_{n-1}] = \cdots = \mathbb{E}[Z_1]$$

Proof.

$$\mathbb{E}[Z_n] = \mathbb{E}\big[\mathbb{E}[Z_n | X_1, \ldots, X_{n-1}]\big] = \mathbb{E}[Z_{n-1}]$$

■

Tensorization of Variance (Revisited)

Theorem. Suppose $X_1, \dots, X_n$ are independent random variables. Let $Z = f(X_1, \dots, X_n)$. Then

$$\mathrm{Var}[Z] \le \mathbb{E}\left[\sum_{i=1}^{n} \mathrm{Var}_i[Z]\right]$$

where $\mathrm{Var}_i[Z](x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n) \coloneqq \mathrm{Var}[f(x_1, \dots, x_{i-1}, X_i, x_{i+1}, \dots, x_n)]$

Proof.

- For $i \in [n]$, define a new random variable Δ_i:

$$\Delta_i \coloneqq \mathbb{E}[Z | X_1, \dots, X_i] - \mathbb{E}[Z | X_1, \dots, X_{i-1}]$$

- Notice the telescoping sum:

$$\sum_{i=1}^{n} \Delta_i = \sum_{i=1}^{n} (\mathbb{E}[Z | X_1, \dots, X_i] - \mathbb{E}[Z | X_1, \dots, X_{i-1}]) = \mathbb{E}[Z | X_1, \dots, X_n] - \mathbb{E}[Z] = Z - \mathbb{E}[Z]$$

Tensorization of Variance (Revisited)

Theorem. Suppose $X_1, \dots, X_n$ are independent random variables. Let $Z = f(X_1, \dots, X_n)$. Then

$$\mathrm{Var}[Z] \le \mathbb{E}\left[\sum_{i=1}^{n} \mathrm{Var}_i[Z]\right]$$

where $\mathrm{Var}_i[Z](x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n) \coloneqq \mathrm{Var}[f(x_1, \dots, x_{i-1}, X_i, x_{i+1}, \dots, x_n)]$

Proof.

- For $i \in [n]$, define a new random variable Δ_i:

$$\Delta_i \coloneqq \mathbb{E}[Z|X_1, \dots, X_i] - \mathbb{E}[Z|X_1, \dots, X_{i-1}]$$

- Moreover, for $i \in [n]$,

$$\begin{aligned}\mathbb{E}[\Delta_i|X_1, \dots, X_{i-1}] &= \mathbb{E}[\mathbb{E}[Z|X_1, \dots, X_i]|X_1, \dots, X_{i-1}] - \mathbb{E}[\mathbb{E}[Z|X_1, \dots, X_{i-1}]|X_1, \dots, X_{i-1}] \\ &= \mathbb{E}[Z|X_1, \dots, X_{i-1}] - \mathbb{E}[Z|X_1, \dots, X_{i-1}] = 0\end{aligned}$$

- We say $\Delta_1, \dots, \Delta_n$ are martingale difference

Tensorization of Variance (Revisited)

Theorem. Suppose $X_1, \dots, X_n$ are independent random variables. Let $Z = f(X_1, \dots, X_n)$. Then

$$\mathrm{Var}[Z] \le \mathbb{E}\left[\sum_{i=1}^{n} \mathrm{Var}_i[Z]\right]$$

where $\mathrm{Var}_i[Z](x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n) \coloneqq \mathrm{Var}[f(x_1, \dots, x_{i-1}, X_i, x_{i+1}, \dots, x_n)]$

Proof.

- For $i \in [n]$, define a new random variable Δ_i:
$$\Delta_i \coloneqq \mathbb{E}[Z|X_1, \dots, X_i] - \mathbb{E}[Z|X_1, \dots, X_{i-1}]$$
- $\mathbb{E}[Z], \mathbb{E}[Z|X_1], \mathbb{E}[Z|X_1, X_2], \dots, \mathbb{E}[Z|X_1, \dots, X_n]$ is a martingale w.r.t. $X_1, \dots, X_n$ (Doob martingale)
- $\mathbb{E}[\Delta_i|X_1, \dots, X_{i-1}] = \mathbb{E}[\mathbb{E}[Z|X_1, \dots, X_i]|X_1, \dots, X_{i-1}] - \mathbb{E}[Z|X_1, \dots, X_{i-1}] = 0$
- For any $j < i$, $\mathbb{E}[\Delta_i\Delta_j] = \mathbb{E}\left[\mathbb{E}[\Delta_i\Delta_j|X_1, \dots, X_{i-1}]\right] = \mathbb{E}[\mathbb{E}[\Delta_i|X_1, \dots, X_{i-1}]\Delta_j] = 0$

(tower property)

Tensorization of Variance (Revisited)

Theorem. Suppose $X_1, \dots, X_n$ are independent random variables. Let $Z = f(X_1, \dots, X_n)$. Then

$$\mathrm{Var}[Z] \le \mathbb{E}\left[\sum_{i=1}^{n} \mathrm{Var}_i[Z]\right]$$

where $\mathrm{Var}_i[Z](x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n) \coloneqq \mathrm{Var}[f(x_1, \dots, x_{i-1}, X_i, x_{i+1}, \dots, x_n)]$

Proof.

- For $i \in [n]$, define a new random variable Δ_i:

$$\Delta_i \coloneqq \mathbb{E}[Z|X_1, \dots, X_i] - \mathbb{E}[Z|X_1, \dots, X_{i-1}]$$

- $Z - \mathbb{E}[Z] = \sum_{i\in[n]} \Delta_i$
- For any $i \neq j$, $\mathbb{E}[\Delta_i \Delta_j] = 0$

$$\mathrm{Var}[Z] = \mathbb{E}[(Z - \mathbb{E}[Z])^2] = \mathbb{E}\left[\left(\sum_{i\in[n]} \Delta_i\right)^2\right] = \sum_{i\in[n]} \mathbb{E}[\Delta_i^2]$$

Tensorization of Variance (Revisited)

Theorem. Suppose $X_1, \dots, X_n$ are independent random variables. Let $Z = f(X_1, \dots, X_n)$. Then

$$\mathrm{Var}[Z] \le \mathbb{E}\left[\sum_{i=1}^{n} \mathrm{Var}_i[Z]\right]$$

where $\mathrm{Var}_i[Z](x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n) \coloneqq \mathrm{Var}[f(x_1, \dots, x_{i-1}, X_i, x_{i+1}, \dots, x_n)]$

Proof.

- For $i \in [n]$, define a new random variable Δ_i:

$$\Delta_i \coloneqq \mathbb{E}[Z|X_1, \dots, X_i] - \mathbb{E}[Z|X_1, \dots, X_{i-1}]$$

- It remains to show that $\mathbb{E}\left[\Delta_i^2\right] \le \mathbb{E}\left[\mathrm{Var}_i[Z]\right]$ for any $i \in [n]$:

$$\begin{aligned}\mathbb{E}[Z|X_1, \dots, X_{i-1}] &= \mathbb{E}[\mathbb{E}[Z|X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n]|X_1, \dots, X_{i-1}] \\ &= \mathbb{E}[\mathbb{E}[Z|X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n]|X_1, \dots, X_{i-1}, X_i]\end{aligned}$$

- Define $\widetilde{\Delta}_i \coloneqq Z - \mathbb{E}[Z|X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n]$. We have $\mathbb{E}\left[\widetilde{\Delta}_i \middle| X_1, \dots, X_i\right] = \Delta_i$

Tensorization of Variance (Revisited)

Theorem. Suppose $X_1, \dots, X_n$ are independent random variables. Let $Z = f(X_1, \dots, X_n)$. Then

$$\mathrm{Var}[Z] \le \mathbb{E}\left[\sum_{i=1}^{n} \mathrm{Var}_i[Z]\right]$$

where $\mathrm{Var}_i[Z](x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n) \coloneqq \mathrm{Var}[f(x_1, \dots, x_{i-1}, X_i, x_{i+1}, \dots, x_n)]$

Proof.

- For $i \in [n]$, define a new random variable Δ_i:
$$\Delta_i \coloneqq \mathbb{E}[Z | X_1, \dots, X_i] - \mathbb{E}[Z | X_1, \dots, X_{i-1}]$$
- Define $\widetilde{\Delta_i} \coloneqq Z - \mathbb{E}[Z | X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n]$. We have $\mathbb{E}\left[\widetilde{\Delta_i} \middle| X_1, \dots, X_i\right] = \Delta_i$
- Since X_i and $\{X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n\}$ are independent,
$$\mathrm{Var}_i[Z] = \mathbb{E}\left[\widetilde{\Delta_k}^2 \middle| X_1, \dots, X_{k-1}, X_{k+1}, \dots, X_n\right]$$
$$\Longrightarrow \quad \mathbb{E}[\Delta_i^2] = \mathbb{E}\left[\mathbb{E}[\widetilde{\Delta_i} | X_1, \dots, X_i]^2\right] \le \mathbb{E}\left[\mathbb{E}\left[\widetilde{\Delta_i}^2 \middle| X_1, \dots, X_i\right]\right] = \mathbb{E}\left[\widetilde{\Delta_i}^2\right] = \mathbb{E}[\mathrm{Var}_i[Z]]$$

(Jensen)

■

Tensorization of Variance (Revisited)

Theorem. Suppose $X_1, \dots, X_n$ are independent random variables. Let $Z = f(X_1, \dots, X_n)$. Then

$$\mathrm{Var}[Z] \le \mathbb{E}\left[\sum_{i=1}^{n} \mathrm{Var}_i[Z]\right]$$

where $\mathrm{Var}_i[Z](x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n) \coloneqq \mathrm{Var}[f(x_1, \dots, x_{i-1}, X_i, x_{i+1}, \dots, x_n)]$

The key idea of the proof is to decompose

$$f(X_1, \dots, X_n) - \mathbb{E}[f(X_1, \dots, X_n)] = \sum_{i=1}^{n} \Delta_i$$

And using the martingale difference property, $\mathrm{Var}[f] = \sum_{i=1}^{n} \mathbb{E}\left[\Delta_i^2\right]$

Today's Lecture

- Tensorization of Variance (Revisited)
- Azuma-Hoeffding Inequality
- Applications
  - Pattern Matching
  - Learning Theory and Glivenko-Cantelli Theorem

Azuma-Hoeffding Inequality

Let $\Delta_1, \Delta_2, \dots, \Delta_n$ be martingale differences and $a_i \le \Delta_i \le b_i$ for any $i \in [n]$. Then

$$\Pr\left[\left|\sum_{i=1}^{n} \Delta_i\right| \ge t\right] \le 2\exp\left(-\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right) \quad \forall t > 0$$

- If we take $\Delta_i = X_i - \mathbb{E}[X_i]$ for independent random variables $X_1, \dots, X_n$ (Think: why are they martingale differences?)
- We recover the Hoeffding inequality:

$$\Pr\left[\left|\sum_{i=1}^{n} (X_i - \mathbb{E}[X_i])\right| \ge t\right] \le 2\exp\left(-\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)$$

Azuma-Hoeffding Inequality

Let $\Delta_1, \Delta_2, \dots, \Delta_n$ be martingale differences and $a_i \le \Delta_i \le b_i$ for any $i \in [n]$. Then

$$\Pr\left[\left|\sum_{i=1}^{n} \Delta_i\right| \ge t\right] \le 2\exp\left(-\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right) \qquad \forall t > 0$$

Proof.

- Let $Z \coloneqq \sum_{i=1}^{n} \Delta_i$. Then $\mathbb{E}[Z] = 0$ and

$$\Pr[Z \ge t] \le \exp\left(\inf_{\theta \ge 0} -\theta t + \psi(\theta)\right)$$

- We just need to bound the log-MGF:

$$\psi(\theta) \coloneqq \log \mathbb{E}\left[e^{\theta Z}\right] = \log \mathbb{E}\left[e^{\theta \sum_{i=1}^{n} \Delta_i}\right]$$

- This is the only step that is different from Hoeffding inequality's proof

Azuma-Hoeffding Inequality

Let $\Delta_1, \Delta_2, \dots, \Delta_n$ be martingale differences and $a_i \le \Delta_i \le b_i$ for any $i \in [n]$. Then

$$\Pr\left[\left|\sum_{i=1}^{n} \Delta_i\right| \ge t\right] \le 2\exp\left(-\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right) \qquad \forall t > 0$$

Proof.

- We just need to bound the log-MGF:

$$\psi(\theta) \coloneqq \log \mathbb{E}\left[e^{\theta Z}\right] = \log \mathbb{E}\left[e^{\theta \sum_{i=1}^{n} \Delta_i}\right]$$

- Suppose $\mathbb{E}[\Delta_i | X_1, \dots, X_{i-1}] = 0$ for any $i \in [n]$, i.e., $\Delta_1, \dots, \Delta_n$ are martingale differences w.r.t. $X_1, \dots, X_n$
- By the tower property,

$$\mathbb{E}\left[e^{\theta \sum_{i=1}^{n} \Delta_i}\right] = \mathbb{E}\left[e^{\theta \sum_{i=1}^{n-1} \Delta_i} \mathbb{E}\left[e^{\theta \Delta_n} \middle| X_1, \dots, X_{n-1}\right]\right]$$

Azuma-Hoeffding Inequality

Let $\Delta_1, \Delta_2, \dots, \Delta_n$ be martingale differences and $a_i \le \Delta_i \le b_i$ for any $i \in [n]$. Then

$$\Pr\left[\left|\sum_{i=1}^{n} \Delta_i\right| \ge t\right] \le 2\exp\left(-\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right) \qquad \forall t > 0$$

Proof.

- By the tower property,

$$\mathbb{E}\left[e^{\theta \sum_{i=1}^{n} \Delta_i}\right] = \mathbb{E}\left[e^{\theta \sum_{i=1}^{n-1} \Delta_i} \mathbb{E}\left[e^{\theta \Delta_n} \middle| X_1, \dots, X_{n-1}\right]\right]$$

- Using the same argument as in the Hoeffding inequality's proof,

$$\mathbb{E}\left[e^{\theta \Delta_n} \middle| X_1, \dots, X_{n-1}\right] \le e^{(b_n - a_n)^2/8}$$

$$\mathbb{E}\left[e^{\theta \sum_{i=1}^{n} \Delta_i}\right] \le e^{(b_n - a_n)^2/8} \mathbb{E}\left[e^{\theta \sum_{i=1}^{n-1} \Delta_i}\right] \le \cdots \le e^{\sum_{i=1}^{n}(b_n - a_n)^2/8}$$

Azuma-Hoeffding Inequality

Let $\Delta_1, \Delta_2, \dots, \Delta_n$ be martingale differences and $a_i \leq \Delta_i \leq b_i$ for any $i \in [n]$. Then

$$\Pr\left[\left|\sum_{i=1}^{n} \Delta_i\right| \geq t\right] \leq 2\exp\left(-\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right) \qquad \forall t > 0$$

Proof.

$$\mathbb{E}\left[e^{\theta \sum_{i=1}^{n} \Delta_i}\right] \leq e^{(b_n - a_n)^2/8}\,\mathbb{E}\left[e^{\theta \sum_{i=1}^{n-1} \Delta_i}\right] \leq \cdots \leq e^{\sum_{i=1}^{n}(b_n - a_n)^2/8}$$

- Thus, $\psi(\theta) = \log \mathbb{E}\left[e^{\theta \sum_{i=1}^{n} \Delta_i}\right] \leq \sum_{i=1}^{n}(b_n - a_n)^2/8$

■

This result can be generalized to case when a_i, b_i are random variables that may depend on $X_1, \dots, X_{i-1}$, and $\Pr[a_i \leq \Delta_i \leq b_i] = 1$ (i.e., $a_i \leq \Delta_i \leq b_i$ almost surely or *a.s.*).

Azuma-Hoeffding Inequality

Let $\Delta_1, \Delta_2, \ldots, \Delta_n$ be martingale differences and $A_i \leq \Delta_i \leq B_i$ *a.s.* for any $i \in [n]$. Then

$$\Pr\left[\left|\sum_{i=1}^{n} \Delta_i\right| \geq t\right] \leq 2\exp\left(-\frac{2t^2}{\sum_{i=1}^{n}\|B_i - A_i\|_\infty^2}\right) \qquad \forall t > 0$$

where $\|B_i - A_i\|_\infty \coloneqq \inf\{c \geq 0 : \Pr[|B_i - A_i| \leq c] = 1\}$

Bounded Differences

Recall that

Corollary. For $Z = f(X_1, \dots, X_n)$, define the i-th discrete partial derivative as:

$$(D_i f)(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n) := \sup_{z \in \mathrm{supp}(X_i)} f(x_1, \dots, x_{i-1}, z, x_{i+1}, \dots, x_n) - \inf_{z \in \mathrm{supp}(X_i)} f(x_1, \dots, x_{i-1}, z, x_{i+1}, \dots, x_n)$$

Then,

$$\mathrm{Var}[Z] \le \frac{1}{4} \sum_{i=1}^{n} \mathbb{E}[(D_i f)^2]$$

We say f satisfies the bounded differences property if there exist $c_1, c_2, \dots, c_n \in \mathbb{R}$ such that

$$\|(D_i f)(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n)\|_\infty \le c_i \qquad \forall i \in [n]$$

McDiarmid Inequality

Let $X_1, \dots, X_n$ be independent random variables and $f(x_1, \dots, x_n)$ be such that satisfies the bounded differences property with $c_1, \dots, c_n$. Then

$$\Pr[|f(X_1, \dots, X_n) - \mathbb{E}[f(X_1, \dots, X_n)]| \geq t] \leq 2 \exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right)$$

Proof.

- We still use the decomposition: $f - \mathbb{E}[f] = \sum_{i=1}^n \Delta_i$, where $\Delta_1, \dots, \Delta_n$ are martingale differences:
$$\Delta_i \coloneqq \mathbb{E}[f(X_1, \dots, X_n)|X_1, \dots, X_i] - \mathbb{E}[f(X_1, \dots, X_n)|X_1, \dots, X_{i-1}]$$
- We need to find random variables A_i, B_i such that $A_i \leq \Delta_i \leq B_i$

$$A_i \coloneqq \mathbb{E}\left[\inf_z f(X_1, \dots, X_{i-1}, z, X_{i+1}, \dots, X_n) \,\middle|\, X_1, \dots, X_{i-1}\right] - \mathbb{E}[f(X_1, \dots, X_n)|X_1, \dots, X_{i-1}]$$
$$B_i \coloneqq \mathbb{E}\left[\sup_z f(X_1, \dots, X_{i-1}, z, X_{i+1}, \dots, X_n) \,\middle|\, X_1, \dots, X_{i-1}\right] - \mathbb{E}[f(X_1, \dots, X_n)|X_1, \dots, X_{i-1}]$$

McDiarmid Inequality

Let $X_1, \dots, X_n$ be independent random variables and $f(x_1, \dots, x_n)$ be such that satisfies the bounded differences property with $c_1, \dots, c_n$. Then

$$\Pr[|f(X_1, \dots, X_n) - \mathbb{E}[f(X_1, \dots, X_n)]| \geq t] \leq 2\exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right)$$

Proof.

$$\begin{aligned}\Delta_i - A_i &= \mathbb{E}[f(X_1, \dots, X_n)|X_1, \dots, X_i] - \mathbb{E}\left[\inf_z f(X_1, \dots, X_{i-1}, z, X_{i+1}, \dots, X_n)\,\middle|X_1, \dots, X_{i-1}\right] \\ &= \mathbb{E}\left[f(X_1, \dots, X_n) - \inf_z f(X_1, \dots, X_{i-1}, z, X_{i+1}, \dots, X_n)\,\middle|X_1, \dots, X_i\right] \\ &\geq 0\end{aligned}$$

- Then, we have

$$|B_i - A_i| = \mathbb{E}[|(D_i f)(X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n)||X_1, \dots, X_{i-1}] \leq c_i$$

■

- Then we complete the proof by Azuma-Hoeffding inequality

Today's Lecture

- Tensorization of Variance (Revisited)
- Azuma-Hoeffding Inequality
- Applications
  - Pattern Matching
  - Learning Theory and Glivenko-Cantelli Theorem

Application 1: Pattern Matching

- Let $X_1, \dots, X_n \in \Sigma$ be a sequence of tokens generated uniformly at random (a trivial language model)
- Let $A = (a_1, \dots, a_k) \in \Sigma^k$ be a fixed length-k token sequence
- Let Z be the number of occurrences of A

➢ What is the expectation of Z?

$$\mathbb{E}[Z] = (n - k + 1) \cdot |\Sigma|^{-k}$$

➢ What is $\Pr[|Z - \mathbb{E}[Z]|]$?

- Consider the martingale differences:

$$\Delta_i \coloneqq \mathbb{E}[Z|X_1, \dots, X_i] - \mathbb{E}[Z|X_1, \dots, X_{i-1}]$$

- Check by yourself that $|\Delta_i| \le k$
- Azuma implies that

$$\Pr[|Z - \mathbb{E}[Z]| \ge t] \le 2e^{-t^2/(2nk^2)}$$

Learning Theory Basics

- For a function $f \in \mathcal{F}$, the empirical risk (with iid data samples $\{(x_i, y_i)\}_{i\in[n]}$) is

$$\hat{R}(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(f(x_i), y_i)$$

- The empirical risk minimizing (ERM) function is $f_n \coloneqq \arg\min_{f\in\mathcal{F}} \hat{R}(f)$

- The true performance (the expected risk) of f is

$$R(f) = \mathbb{E}_{(x,y)\in\mathcal{D}}[\ell(f(x), y)]$$

  and we define $f^\star \coloneqq \arg\min_{f\in\mathcal{F}} R(f)$

- We want to control the excess risk:

$$R(f_n) - R(f_\star) = \underbrace{R(f_n) - \hat{R}(f_n)}_{\text{Uniform laws of large numbers for } \mathcal{F}} + \underbrace{\hat{R}(f_n) - \hat{R}(f^\star)}_{\leq 0 \text{ by ERM}} + \underbrace{\hat{R}(f^\star) - R(f^\star)}_{\text{LLN for } f^\star}$$

Learning Theory Basics

- For a function $f \in \mathcal{F}$, the empirical risk (with iid data samples $\{(x_i, y_i)\}_{i\in[n]}$) is

$$\hat{R}(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(f(x_i), y_i)$$

- The empirical risk minimizing (ERM) function is $f_n \coloneqq \arg\min_{f\in\mathcal{F}} \hat{R}(f)$
- The true performance (the expected risk) of f is

$$R(f) = \mathbb{E}_{(x,y)\in\mathcal{D}}[\ell(f(x), y)]$$

  and we define $f^\star \coloneqq \arg\min_{f\in\mathcal{F}} R(f)$

$$\hat{R}(f^\star) - R(f^\star) = \sum_{i=1}^{n} \frac{1}{n}\ell(f^\star(x_i), y_i) - \mathbb{E}[\ell(f^\star(x), y)]$$

- For a bounded loss function ℓ, Hoeffding's inequality implies that this error converges to 0 w.h.p.

Learning Theory Basics

- For a function $f \in \mathcal{F}$, the empirical risk (with iid data samples $\{(x_i, y_i)\}_{i\in[n]}$) is

$$\hat{R}(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(f(x_i), y_i)$$

- The empirical risk minimizing (ERM) function is $f_n \coloneqq \arg\min_{f\in\mathcal{F}} \hat{R}(f)$
- The true performance (the expected risk) of f is

$$R(f) = \mathbb{E}_{(x,y)\in\mathcal{D}}[\ell(f(x), y)]$$

  and we define $f^\star \coloneqq \arg\min_{f\in\mathcal{F}} R(f)$
- The first term $R(f_n) - \hat{R}(f_n)$ is more interesting. f_n is a random function depending on $\{(x_i, y_i)\}_{i\in[n]}$
- We can upper bound it by $R(f_n) - \hat{R}(f_n) \leq \sup_{f\in F} \left|R(f) - \hat{R}(f)\right|$
- The uniform laws of large numbers provide an upper bound for the excess risk for all functions

Glivenko-Cantelli Theorem

Let $X_1, X_2, \ldots$ be iid random variables with the cumulative distribution function (CDF) $F(x)$

Define the empirical distribution function for $X_1, \ldots, X_n$ as

$$F_n(x) := \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}[X_i \le x]$$

Then, $\|F_n - F\|_\infty = \sup_x |F_n(x) - F(x)| \xrightarrow{a.s.} 0$

- Let P be the distribution of each X_i, and P_n be the empirical distribution (with CDF F_n)
- GC theorem implies that $\sup_x |F_n(x) - F(x)| = \sup_x \left| \Pr_{X\sim P_n}[X \le x] - \Pr_{X\sim P}[X \le x] \right| \xrightarrow{a.s.} 0$
- Define a function class $G := \{\mathbf{1}[x \le t] : t \in \mathbb{R}\}$
- Then, GC theorem $\Leftrightarrow \sup_{g\in G} \left| \mathbb{E}_{P_n}[g] - \mathbb{E}_P[g] \right| =: \|P_n - P\|_G \xrightarrow{a.s.} 0$

Glivenko-Cantelli Theorem

Let $X_1, X_2, \ldots$ be iid random variables with the cumulative distribution function (CDF) $F(x)$

Define the empirical distribution function for $X_1, \ldots, X_n$ as

$$F_n(x) \coloneqq \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}[X_i \le x]$$

Then, $\|F_n - F\|_\infty = \sup_x |F_n(x) - F(x)| \xrightarrow{a.s.} 0$

Proof (Key ideas).

1. Concentration: $\|P_n - P\|_G \approx \mathbb{E}[\|P_n - P\|_G]$ w.h.p.
2. Symmetrization: $\mathbb{E}[\|P_n - P\|_G] \le 2\mathbb{E}[\|R_n\|_G]$ where $\mathbb{E}_{R_n}[g] \coloneqq (1/n)\sum_{i=1}^{n} \epsilon_i g(X_i)$ (Rademacher process)
3. Restriction: G restricted to a finite-sized set to bound the Rademacher averages

Glivenko-Cantelli Theorem: Concentration

$$\|P_n - P\|_G = \sup_{g \in G} \left| \mathbb{E}_{P_n}[g(X)] - \mathbb{E}_P[g(X)] \right| = \sup_{g \in G} \left| \sum_{i=1}^{n} \frac{1}{n} \mathbf{1}[X_i \leq t] - \mathbb{E}_P[g(X)] \right|$$

- $\|P_n - P\|_G$ is a function of $X_1, \ldots, X_n$
- It has the bounded differences property:

$$\sup_{z} \|P_n - P\|_G(X_1, \ldots, X_{i-1}, z, X_{i+1}, \ldots, X_n) - \inf_{z} \|P_n - P\|_G(X_1, \ldots, X_{i-1}, z, X_{i+1}, \ldots, X_n) \leq \frac{1}{n}$$

- McDiarmid inequality: with probability at least $1 - \exp(-2\epsilon^2 n)$,

$$\|P_n - P\|_G \leq \mathbb{E}[\|P_n - P\|_G] + \epsilon$$

Glivenko-Cantelli Theorem: Symmetrization

- Note that for iid samples $X_1', \dots, X_n'$, $\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} g(X_i')\right] = \mathbb{E}_P[g]$
- Thus, we can introduce another n iid samples $X_1', \dots, X_n'$, and get that

$$\mathbb{E}[\|P_n - P\|_G] = \mathbb{E}_{X_i}\left[\sup_{g\in G}\left|\mathbb{E}_{X_i'}\left[\frac{1}{n}\sum_{i=1}^{n}\left(g(X_i) - g(X_i')\right)\right]\right|\right]$$

$$\leq \mathbb{E}_{X_i}\mathbb{E}_{X_i'}\left[\sup_{g\in G}\left|\frac{1}{n}\sum_{i=1}^{n}\left(g(X_i) - g(X_i')\right)\right|\right]$$

$$= \mathbb{E}[\|P_n - P_n'\|_G]$$

- The second step follows from Jensen inequality and the fact that $\sup|\cdot|$ is convex

Glivenko-Cantelli Theorem: Symmetrization

- Since $\{X_i, X_i'\}$ are iid, for any $\epsilon_i \in \{-1,1\}$,

$$\mathbb{E}\left[\sup_{g\in G}\left|\frac{1}{n}\sum_{i=1}^{n}\left(g(X_i)-g(X_i')\right)\right|\right]=\mathbb{E}\left[\sup_{g\in G}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\left(g(X_i)-g(X_i')\right)\right|\right]$$

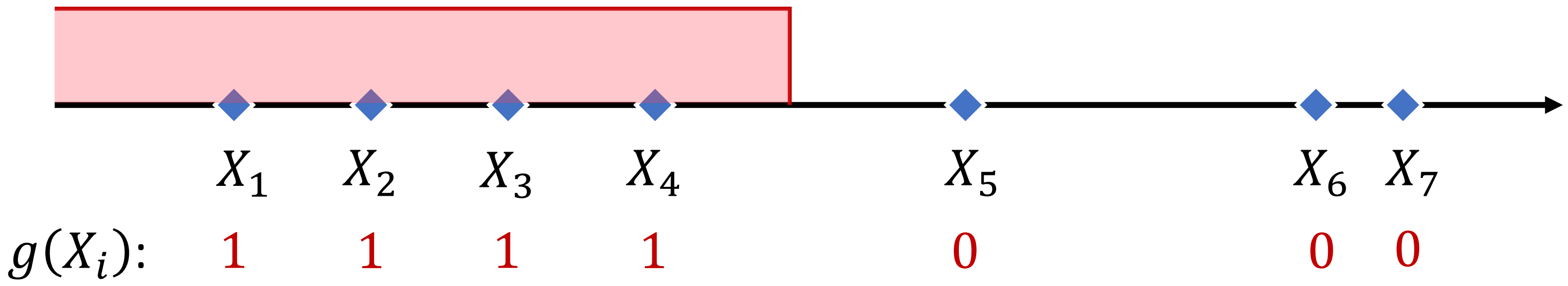
- The equality still holds if we take the expectation over $\epsilon_i \sim_{iid} \{-1,1\}$ uniformly at random

$$\begin{aligned}\mathbb{E}[\|P_n-P\|_G] &= \mathbb{E}_{X_i,X_i',\epsilon_i}\left[\sup_{g\in G}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\left(g(X_i)-g(X_i')\right)\right|\right]\\ &\le \mathbb{E}\left[\sup_{g\in G}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i g(X_i)\right|\right]+\mathbb{E}\left[\sup_{g\in G}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i g(X_i')\right|\right]\\ &= 2\mathbb{E}\left[\sup_{g\in G}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i g(X_i)\right|\right] =: 2\mathbb{E}[\|R_n\|_G]\end{aligned}$$

Glivenko-Cantelli Theorem: Restriction

$$\mathbb{E}[\|R_n\|_G] = \mathbb{E}\left[\sup_{g \in G}\left|\frac{1}{n}\sum_{i=1}^{n} \epsilon_i g(X_i)\right|\right]$$

- $G = \{\mathbf{1}[x \leq t] : t \in \mathbb{R}\}$ has ∞-many elements
- For any fixed $X_1, \dots, X_n \in \mathbb{R}$, the restriction $G(X_1, \dots, X_n) = \{\{g(X_1), \dots, g(X_n)\} : g \in G\}$ has only $n + 1$ elements!



Glivenko-Cantelli Theorem: Restriction

$$\mathbb{E}[\|R_n\|_G] = \mathbb{E}\left[\sup_{g \in G} \left|\frac{1}{n}\sum_{i=1}^{n} \epsilon_i g(X_i)\right|\right]$$

- $G = \{\mathbf{1}[x \leq t] : t \in \mathbb{R}\}$ has ∞-many elements
- For any fixed $X_1, \dots, X_n \in \mathbb{R}$, the restriction $G(X_1, \dots, X_n) = \left\{\{g(X_1), \dots, g(X_n)\} : g \in G\right\}$ has only $n + 1$ elements!

Lemma (Rademacher averages). For a finite subset $A \subseteq \mathbb{R}^n$ and $\sigma^2 \coloneqq \max_{a \in A} \|a\|_2^2 / n$,

$$\mathbb{E}_{\epsilon_i \sim \{\pm 1\}}\left[\sup_{a \in A} \frac{1}{n}\sum_{i=1}^{n} \epsilon_i a_i\right] \leq \sqrt{\frac{2\sigma^2 \log|A|}{n}}$$

$$\mathbb{E}\left[\sup_{a \in A} \frac{1}{n}\left|\sum_{i=1}^{n} \epsilon_i a_i\right|\right] = \mathbb{E}\left[\sup_{a \in A \cup (-A)} \frac{1}{n}\sum_{i=1}^{n} \epsilon_i a_i\right] \leq \sqrt{\frac{2\sigma^2 \log(2|A|)}{n}}$$

Glivenko-Cantelli Theorem: Restriction

$$\mathbb{E}[\|R_n\|_G] = \mathbb{E}\left[\sup_{g\in G}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i g(X_i)\right|\right]$$

- $G = \{\mathbf{1}[x \le t] : t \in \mathbb{R}\}$ has ∞-many elements
- For any fixed $X_1, \dots, X_n \in \mathbb{R}$, the restriction $G(X_1, \dots, X_n) = \{\{g(X_1), \dots, g(X_n)\} : g \in G\}$ has only $n + 1$ elements!

Lemma (Rademacher averages). For a finite subset $A \subseteq \mathbb{R}^n$ and $\sigma^2 \coloneqq \max_{a\in A}\|a\|_2^2/n$,

$$\mathbb{E}\left[\sup_{a\in A}\frac{1}{n}\left|\sum_{i=1}^{n}\epsilon_i a_i\right|\right] \le \sqrt{\frac{2\sigma^2\log(2|A|)}{n}}$$

- In our case, $|A| \le n + 1$ and $\sigma^2 \le n/n = 1$:

$$\mathbb{E}[\|R_n\|_G] \le \sqrt{\frac{2\log\big(2(n+1)\big)}{n}}$$

Glivenko-Cantelli Theorem: Putting Together

- Concentration:

$$\Pr[\|P_n - P\|_G \le \mathbb{E}[\|P_n - P\|_G] + \epsilon] \ge 1 - \exp(-2\epsilon^2 n)$$

- Symmetrization:

$$\mathbb{E}[\|P_n - P\|_G] \le 2\mathbb{E}[\|R_n\|_G]$$

- Restriction:

$$\|R_n\|_G \le \sqrt{\frac{2\log\big(2(n+1)\big)}{n}}$$

- Therefore,

$$\Pr\left[\|P_n - P\|_G \le \sqrt{\frac{8\log\big(2(n+1)\big)}{n}} + \epsilon\right] \ge 1 - e^{-2\epsilon^2 n}$$

■

Proof of Rademacher Averages Lemma

Lemma (Rademacher averages)**.** For a finite subset $A \subseteq \mathbb{R}^n$ and $\sigma^2 \coloneqq \max_{a \in A} \|a\|_2^2 / n$,

$$\mathbb{E}_{\epsilon_i \sim \{\pm 1\}}\left[\sup_{a \in A} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i a_i\right] \leq \sqrt{\frac{2\sigma^2 \log|A|}{n}}$$

Proof.

- Consider the MGF:

$$\exp\left(\theta \mathbb{E}\left[\sup_{a \in A} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i a_i\right]\right) \leq \mathbb{E}\left[\exp\left(\theta \sup_{a \in A} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i a_i\right)\right] = \mathbb{E}\left[\sup_{a \in A} \exp\left(\theta \frac{1}{n} \sum_{i=1}^{n} \epsilon_i a_i\right)\right]$$

$$\leq \sum_{a \in A} \mathbb{E}\left[\exp\left(\theta \frac{1}{n} \sum_{i=1}^{n} \epsilon_i a_i\right)\right] = \sum_{a \in A} \prod_{i=1}^{n} \mathbb{E}\left[\exp\left(\frac{\theta a_i}{n} \epsilon_i\right)\right]$$

(Hoeffding)

$$\leq \sum_{a \in A} \prod_{i=1}^{n} \exp\left(\frac{\theta^2 a_i^2}{2n^2}\right) = \sum_{a \in A} \exp\left(\frac{\theta^2 \|a\|_2^2}{2n^2}\right) \leq |A| \exp\left(\frac{\theta^2 \sigma^2}{2n}\right)$$

Proof of Rademacher Averages Lemma

Lemma (Rademacher averages)**.** For a finite subset $A \subseteq \mathbb{R}^n$ and $\sigma^2 \coloneqq \max_{a\in A} \|a\|_2^2/n$,

$$\mathbb{E}_{\epsilon_i \sim \{\pm 1\}}\left[\sup_{a\in A} \frac{1}{n}\sum_{i=1}^{n} \epsilon_i a_i\right] \le \sqrt{\frac{2\sigma^2 \log|A|}{n}}$$

Proof.

- Consider the MGF:

$$\exp\left(\theta \mathbb{E}\left[\sup_{a\in A} \frac{1}{n}\sum_{i=1}^{n} \epsilon_i a_i\right]\right) \le |A| \exp\left(\frac{\theta^2\sigma^2}{2n}\right)$$

- Thus, we have

$$\mathbb{E}\left[\sup_{a\in A} \frac{1}{n}\sum_{i=1}^{n} \epsilon_i a_i\right] \le \frac{\log|A|}{\theta} + \frac{\theta\sigma^2}{2n} = \sqrt{\frac{2\sigma^2 \log|A|}{n}} \quad \text{with } \theta \coloneqq \sqrt{2n \log|A|}/\sigma$$

■

Glivenko-Cantelli Theorem

Let $X_1, X_2, \ldots$ be iid random variables. Define the empirical distribution P_n by its CDF:

$$F_n(x) \coloneqq \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}[X_i \le x]$$

Then, for the function family $G = \{\mathbf{1}[x \le t] : t \in \mathbb{R}\}$, we have

$$\sup_{g \in G} \left| \mathbb{E}_{P_n}[g] - \mathbb{E}_P(g) \right| \xrightarrow{a.s.} 0$$

- Generalizing the GC theorem to GC class (the function class that satisfies the uniform convergence)
- GC class is connected to the Vapnik-Chervonenkis (VC) dimension

The Fundamental Theorem of Statistical Learning

Let $\mathcal{C}$ be a concept class of functions from a domain $\mathcal{X}$ to $\{-1,1\}$, and let the loss function be the 0-1 loss (i.e., $\mathbf{1}[f(x) \neq y]$). Then the following are equivalent:

1. $\mathcal{C}$ has the uniform convergence property
2. $\mathcal{C}$ is (agnostic) PAC learnable
3. $\mathcal{C}$ is (realizable) PAC learnable
4. $\mathcal{C}$ has finite VC dimension
5. $\mathcal{C}$ is learnable by an ERM algorithm

Covered in CS 578 - Statistical Machine Learning by Anuran Makur